Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 6654096, 11 pages
https://doi.org/10.1155/2021/6654096

Research Article

Modeling of Merging Decision during Execution Period Based on Random Forest

Gen Li[1], **Jianxiao Ma**[1] **and Qiangru Shen**[2]

[1]*College of Automobile and Traffic Engineering, Nanjing Forestry University, Nanjing 210037, China*
[2]*School of Transportation and Civil Engineering, Nantong University, Nantong 226019, China*

Correspondence should be addressed to Gen Li; ligen@njfu.edu.cn

Received 9 November 2020; Revised 21 January 2021; Accepted 25 January 2021; Published 3 February 2021

Academic Editor: Zhibin Li

This study aims to investigate the key feature variables and build an accurate decision model for merging behavior during the execution period by using a data-driven method called random forest (RF). To comprehensively explore the feature variables during merging execution period, nineteen candidate variables including speeds, relative speeds, gaps, time-to-collisions (TTCs), and locations are extracted from a dataset including 375 noise-filtered vehicle trajectories. After the variable selection process, an RF model with 9 key feature variables is finally built. Results show that the gap between the merging vehicle and its putative following vehicle and the ration of this gap to the total accepted gap are the two most important feature variables. It is because merging vehicle drivers can easily observe the putative leading vehicles and control the relative speeds and positions to the putative leading vehicles and they tend to leave more space for their putative following vehicles. Relative speed between the merging vehicle and its following vehicle in the auxiliary lane is the only variable related to the vehicles in the auxiliary lane, which means merging vehicles mainly focus on the traffic condition in the adjacent main lane. Evaluation of the performance in comparison with the state-of-the-art method reveals that the proposed method can obtain much more accurate results in both training and testing datasets, which means RF is practical for predicting the merging decision behavior during execution period and has better transferability.

1. Introduction

As a basic driving task, lane changing has drawn great attention recently. Lane changing behavior was considered to be an important reason for traffic oscillations and accidents [1–4]. It was estimated that lane change crashes account for 4 to 10% of all crashes in the US [5]. Lane-changing behavior is complicated and risky because it is influenced by vehicles in both the current lane and the target lane. Several factors such as velocities and gaps should be taken into account during the lane changing process.

Luckily, with the rapid development of communication technology, driving assistance systems have been developed to help drivers to make safer decisions [6, 7]. Lane-changing decision assistance is one of the key functions of driving assistance systems. It can help drivers make safer decisions to start a lane change. Through the Vehicular Ad-hoc Network (VANET), vehicles can communicate with the surrounding vehicles and roadside unites [8–10]. The lane-changing decision assistance systems can well deal with the situation of discretionary lane-changing by using the data from surrounding vehicles and roadside unites. However, for merging areas on freeway, the judgment rules might be not applicable [11]. In merging areas, drivers need to change to the adjacent main lane within the limited distance, which may result in traffic congestions and even breakdowns [12–17].

As a sequential decision process, the whole merging process can be simplified as a sequential two-step model (gap searching and merging execution) or a three-step model (gap searching, merging position searching, and merging execution) [18–21]. However, most previous studies focused on the gap searching process but neglected the merging execution period. Several seconds are needed to execute the merging behavior and the traffic condition may change dynamically during the whole

merging execution period. The ignorance of the merging execution process would lead to reduction of accuracy of traffic simulation and autonomous driving. Thus, there is a critical need to model the merging decision behavior during the execution period. During the merging execution period, the merging vehicles have interactions with putative leading (PL) and putative following (PF) vehicles in the adjacent main lane and the leading (L) and following (F) vehicles in the auxiliary lane. Various influencing factors might be considered for merging decision and should be analyzed in depth. However, previous studies [17] showed that there is multicollinearity between the variables. It was pointed by Balal et al. [22] that most of the lane changing related variables are highly correlated, implying that only a few representative or key variables might be sufficient to describe the interactions of vehicles. However, the selection of key variables is not an easy work. Therefore, the variable selection process should be conducted before building parametric models such as logit model. Improper selection of the key variables might make the performance of the model deteriorate too seriously to be applied to merging assistance systems.

Recently, data mining techniques have received a lot of attention in transportation fields due to their ability to deal with the large-scale data. Some of them can naturally overcome the multicollinearity problem and make full use of the training data. Thus, this study tried use a famous machine learning technique, random forest (RF), to model the merging decision behavior during execution period. It can not only produce more accurate prediction results but also excavate the hidden information among the data. More importantly, RF can effectively select the key variables. The main contribution can be summarized as follows: first, this study gives a comprehensive analysis of the influencing variables of merging decision. Second, the proposed RF method can accurately predict the merging decision during execution period, which can improve the safety and comfort level of driving assistance system if it could be incorporated into lane changing assistance system. Third, a key feature selection process is conducted to investigate the influencing factors. These contributions can not only help understand the diverse influences of different variables on the merging decision but also shed new insights for driver assistance systems and autonomous driving.

The remainder of the paper is organized as follows. Section 2 will provide a state-of-the-art review on the existing studies followed by section 3, which gives the methodology to build a RF model. Section 4 describes the NGSIM data used in this paper and comprehensively analyzes the influencing variables. Results and discussions are presented in section 5. Finally, the concluding remarks are presented in section 6.

2. Literature Review

Predicting merging decision has always been one of the focuses of transportation researches. A great number of models have been developed based on different theories. The first comprehensive lane changing framework was developed by Gipps [23] based on gap acceptance theory. Then, similar frameworks were adopted in other studies [24–27]. However, the gap acceptance theory has been criticized that it cannot reflect the real behavior of drivers. To overcome the deficiency, logistic and logit models were introduced by some researchers [15, 28, 29]. To account for the heterogeneity among drivers, mixed models were proposed by Weng et al. [30] and Li [31]. Game theory models were also developed to model the merging behavior [32, 33]. However, the prediction accuracy of the parametric models is barely satisfactory and the collinearity of influencing variables makes it difficult for researchers to choose appropriate variables to build accurate models [22].

Recently, data-driven methods, such as classification and regression tree (CART), Bayesian network, and fuzzy logic models, were used in building merging models or lane changing models and achieved promising results [16, 34–38]. CART was applied by Weng et al. [11] to model the merging decision in work zone area during execution period, in which time-to-collision (TTC) was considered as a risky factor. Considering the difference between cars and heavy vehicles, Moridpour et al. [39] presented the lane changing model based on fuzzy logic for heavy vehicles. A cooperative merging strategy was developed by Xu et al. [40] for vehicles with V2V and V2I networks, which is applicable to cooperative merging operations under saturated traffic conditions. However, the majority of previous studies separately considered speeds, relative speeds, and gaps as the influencing variables and ignored the interaction of variables. In addition, considering the complexity of merging behavior, a comprehensive analysis of all possible influencing factors should be conducted to better understand the merging decision during execution period.

Previous studies showed that the variables of lane changing behaviour were highly correlated with each other [17, 22, 31]. Thus, selecting some representative or key variables might better describe the interactions of vehicles. However, feature selection has never been an easy work. Feature selection methods can be classified into statistics based methods [41], information theory [42], manifold [43], and rough set [44]. Besides, data-driven methods are also widely used for feature selection [34, 45, 46]. In this study, a popular data-driven method called random forest was applied in this paper to model the merging decision during the execution period. Compared with other models in the literature, the RF has several unique features and advantages. First, it is able to handle multisource heterogeneous data without long-time data processing. Second, as an ensemble machine learning technique based on CART, RF inherits the advantage of CART that can automatically accommodate missing data of independent variables. Third, RF overcomes the deficiency of CART and can automatically resist outliers and is not easy to be affected by small perturbations in the training data. Finally, RF can select the key variables from high dimension data by the importance of all independent variables [45, 47]. RF has been successfully used in traffic prediction and produced promising results [48–51].

3. Methodology

Predicting merging decision can be simplified as a classification problem. Some classical machine learning techniques, such as CART, are very suitable for modeling merging decision. Though CART is efficient and easy-to-use, it is also easy to be affected by small perturbations in the training data [52]. To improve the robustness and generalization capacity of CART, an ensemble learning technique called random forest, which combines the bagging technique, CART, and random subspace method, was proposed by Breiman [45]. RF is an ensemble classifier composed of a group of decision tree classifiers and gets the prediction result by a simple majority vote. The RF model can improve the prediction accuracy of merging decision as well as help connected and autonomous vehicles (CAVs) make safer decisions during merging process. A brief description of random forest is given in this section and detailed fundamentals of mathematics can be referred to Breiman [45].

In RF, bootstrap aggregating (bagging) is the most basic theory. Suppose we have a training dataset (X, Y) with N training samples $\{(X_1, y_1), (X_2, y_2), \ldots, (X_N, y_N)\}$, where $X_i = \{x_i^1, x_i^2, \ldots, x_i^K\}$ and y_i represent the feature vector and the response variable of the sample i, respectively. Through bagging, RF generates B new training sets (X^b, Y^b) by sampling from (X, Y) uniformly and with replacement for N times. By sampling with replacement, some observations may be repeated in each data set (X^b, Y^b) and some may not appear. The probability that each sample in (X^b, Y^b) not selected is $(1 - (1/N))^N$.

Then, we can get

$$\lim_{N \longrightarrow \infty} \left(1 - \frac{1}{N}\right)^N = \frac{1}{e} \approx 0.368. \quad (1)$$

Equation (1) indicates that about 36.8% of the samples are not used in the training process, which is called OOB (Out of Bag) data. These data can be used for validation. Thus, cross-validation or separate test data are not necessary like other machine learning methods. In RF, the OOB error has been proved to be an unbiased estimation of generalization error.

The random subspace method is also used in RF. It can also be called attribute bagging or feature bagging, which means each tree is constructed based on a random subset of the feature variables. This method is designed to reduce the correlation between the trees and improve the generalization accuracy because the RF uses a simple majority vote of all the trees.

Combining the above two methods and CART, the basic steps of RF can be shown in Figure 1 and summarized as follows:

(I) Initiate the algorithm, set $b = 1$.

(II) Use the bootstrap sampling method to obtain a new data set (X^b, Y^b) by random sampling with replacement for N times, and the data that are not sampled will form a set called OOB set.

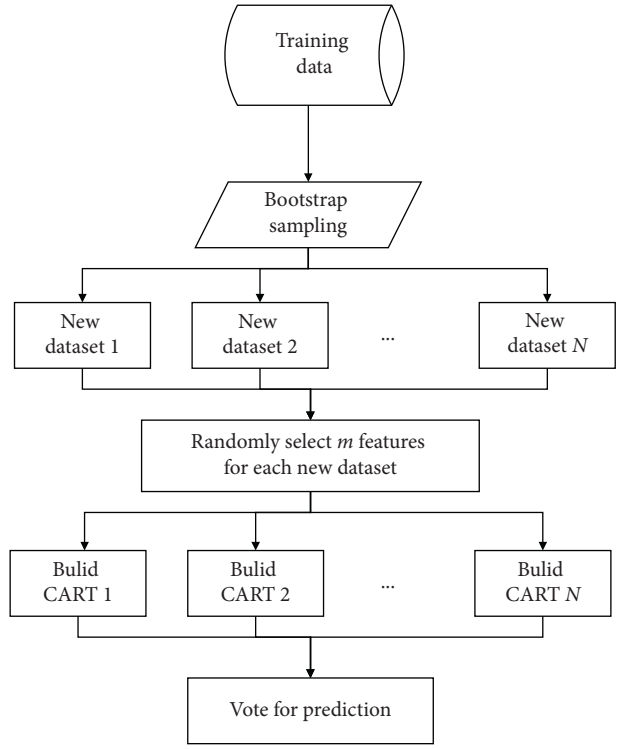

Figure 1: Flow chart of random forest.

(III) Randomly select m feature variables $(m < J)$ and use the selected variables for splitting to train a decision tree T_b based on the new sample set (X^b, Y^b). The decision tree will grow the deepest and is not pruned.

(IV) For $b = 2, \ldots, B$, repeats steps II-III.

The importance of the variables can be sorted by OOB data. RF can screen out important variables in the complex feature variable space, which is conducive to deepen the understanding of the research object. Assuming that the sample subset obtained by bootstrap method is $b = 1, 2, \ldots, B$, the process of using RF to calculate the importance of variable x_j is as follows:

(1) Suppose $b = 1$, and determine the OOB data $L_{b,j}^{\mathrm{OOB}}$.

(2) Use T_b to predict OOB data $L_{b,j}^{\mathrm{OOB}}$, and get the number of accurate predicted samples R_b^{OOB}.

(3) For the feature variable x_j, $j = 1, \ldots, J$, the following calculations are adopted:

(a) Randomly change the variable values x_j in L_b^{OOB} to get a new data set $L_{b,j}^{\mathrm{OOB}}$

(b) Use T_b and L_b^{OOB} for prediction and get the number of correct classification R_b^{OOB}

(c) Calculate the reduction value of classification accuracy, $R_b^{\mathrm{OOB}} - R_{b,j}^{\mathrm{OOB}}$

(4) For $b = 2, \ldots, B$, repeat steps (1–3), and calculate the average value of the reduced value of the classification accuracy to obtain the importance measurement of the variable x_j:

$$D_j = \frac{1}{B}\sum_{b}^{B}\left(R_b^{\mathrm{OOB}} - R_{b,j}^{\mathrm{OOB}}\right). \quad (2)$$

Previous studies have shown that the merging decision could be influenced by a number of highly correlated variables [22, 35]. Thus, the feature selection process must be conducted before building parametric merging decision models. By bagging and random space method, RF can naturally overcome the collinearity of influencing variables. Furthermore, the importance values can be utilized to rank the influencing variables and select the key feature variables through a forward stepwise or backward stepwise elimination process, which will be described in section 5.3.

4. Data Preparation

4.1. Data Description and Processing. In this section, vehicle trajectory data collected by the Federal Highway Administration (FHWA) in the NGSIM project are adopted to verify the proposed RF model. As an open-source dataset, the NGSIM dataset can provide rich and accurate vehicle trajectory data collected on both freeway and urban road [14]. It has been widely used in traffic studies such as traffic flow analysis and driving behavior modeling [18, 37, 53, 54].

Previous studies have shown that the US-101 dataset had the best accuracy and consistency [18, 55]. Thus, this dataset is chosen in this study. Figure 2 shows schematic diagram of data collecting site. One can find that the chosen 640 meters long segment is located between an on-ramp and an off-ramp with five main lanes and one auxiliary lane. Videos were captured from 7:50 a.m. to 8:35 a.m. on June 15, 2005, which was a sunny day. The dataset is updated at a resolution of 10 fps (frames per second) and contains three subsets containing 15 minutes trajectory data [56]. Table 1 shows the aggregate statics of speed and volume for every subset. The coordinates, speed, and acceleration of every vehicle at any instant can be easily obtained from the NGSIM dataset. Previous studies have shown that some random noises existed in the NGSIM data [55, 57]. Filtering and smoothing techniques should be adopted before using. In this study, a data smoothing technique called symmetric exponential moving average filter (sEMA) proposed by Thiemann et al. [57] is applied before further data analysis. In addition, the local coordinates of three subsets are unified to filter the inconsistency of the local coordinates. Detailed steps of data processing can be referred to Li and Sun [17], Li [31], and Li and Cheng [15]. After processing, trajectories of 375 merging vehicle trajectories are extracted from the dataset. All of the vehicles are passenger cars with lengths from 2.5 m to 7.8 m.

4.2. Data Extraction. After selecting the accepted gap, one merging vehicle needs several seconds to find the right time to merge into the adjacent lane and the driver may keep on adjusting the speed and relative position through acceleration deceleration during the execution period. At any time, a merging driver can either choose to continue merging or complete merging as shown in Figure 3. Let y_n^t define the n^{th} merging vehicle's decision at time t. Obviously, y_n^t is a binary variable, shown in the following equation:

$$y_n^t = \begin{cases} 1 \text{ merging vehicle } n \text{ selects to complete merging at time } t \\ 0 \text{ merging vehicle } n \text{ select to continue adjusting at time } t \end{cases}, \quad n = 1, \ldots, N,\ t = 1, \ldots, T_n. \quad (3)$$

Previous studies showed one second is suitable for a driver to make decisions [11, 28, 34, 37]. Thus, we also choose one second in this study. Then, T_n represents the total time to complete merging for vehicle n. Obviously, a merging vehicle can have several observations of $y_n^t = 0$, but only have one observation of $y_n^t = 1$. By extracting the trajectory data of 375 merging vehicles, 1583 observations are obtained in this paper, that is, 375 observations are selecting to merge ($y_n^t = 1$) and 1208 observations are not ($y_n^t = 0$). It means that it takes 3.23 seconds on average for a vehicle to complete merging after making the decision of gap selection.

During the process of merging execution, it has some certain influence on the additional lane and the main lane. At the same time, the merging behavior is also affected by the traffic flow state of the two lanes and the surrounding vehicles. Therefore, the main factors that affect the decision-making of merging vehicles are the speeds, relative speeds, and gaps in the adjacent main lane and the auxiliary lane.

However, previous models considered the above variables separately and ignored the interaction between variables. Some studies showed that the gaps between the merging vehicle and PF vehicle in adjacent main line were linearly related to the total gap during the merging process [20]. Figure 4 shows the scatter plots of the PF gaps and the accepted gaps according to the dataset used in this study. A strong linear relationship can be found in Figure 4. One can also find that the range of the ratio of the PF gap to the accepted gap for $y_n^t = 1$ is rather smaller than that for $y_n^t = 0$, indicating that this ratio might be an important factor for merging decision. Therefore, the ratio of the PF gap to the accepted gap is also considered as the influence variable in this paper.

In addition, a surrogate safety measure combining vehicle speeds, space gap, and time-to-collision (TTC) was also considered, because merging driver needs to control vehicle to avoid rear end accidents with the surrounding vehicles. TTC is defined as

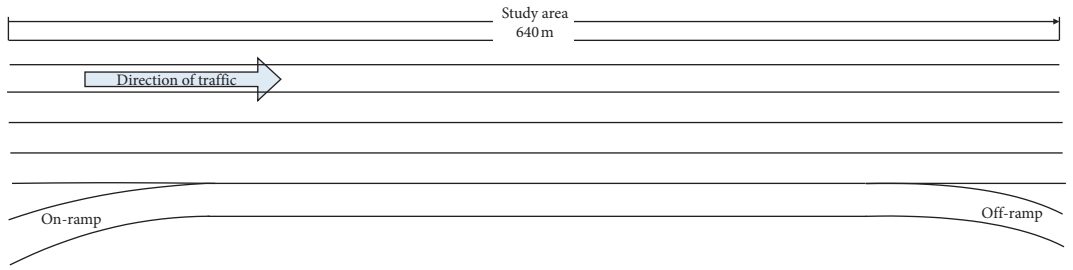

Figure 2: Schematic diagram of U.S. Highway 101.

Table 1: Aggregate statics of three subsets.

Time period	Main lane		Auxiliary lane	
	Volume (vph)	Time mean speed (km/h)	Volume (vph)	Time mean speed (km/h)
7:50 a.m.~8:05 a.m.	8148	44.00	464	63.99
8:05 a.m.~8:20 a.m.	7552	38.80	464	59.26
8:20 a.m.~8:35 a.m.	7108	33.61	496	55.44

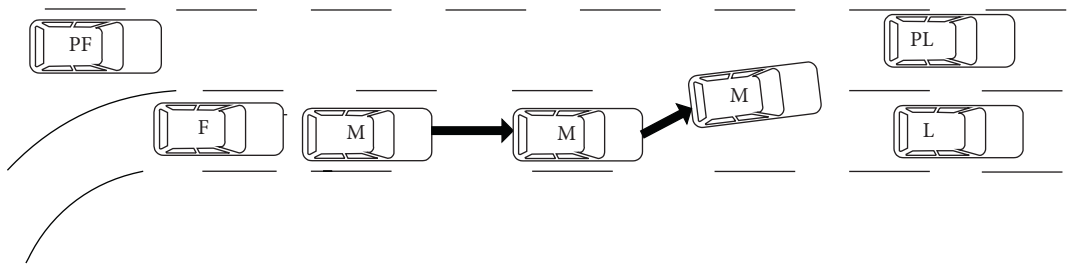

Figure 3: Merging decision during from the start to the end of the execution process.

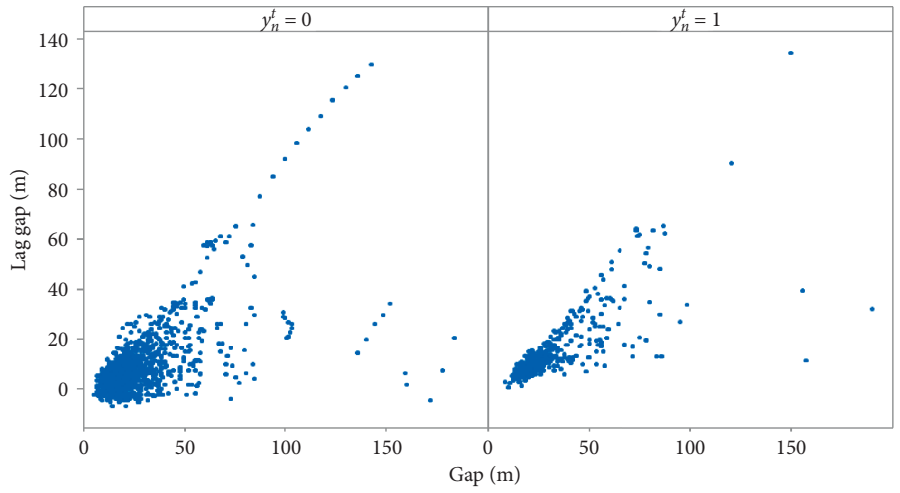

Figure 4: Relationship between the lag gap and the accepted gap in the main line.

$$\text{TTC} = \frac{x_L - x_F - L}{V_F - V_L}, \tag{4}$$

where x_L and x_F are the longitudinal position coordinates of the front bumper of the leading and following vehicle, respectively; V_L and V_F are the speeds of leading and following vehicle, respectively; and L is the length of leading vehicle.

Figure 5 shows the interactions between a merging vehicle and its surrounding vehicles. Table 2 shows the candidate variables and their explanations. It should be pointed out that TTC is negative when the following vehicle moves slower than the leading vehicle, which means that the collision would never occur. In addition, when the speed of the following vehicle is equal to or slightly larger than the

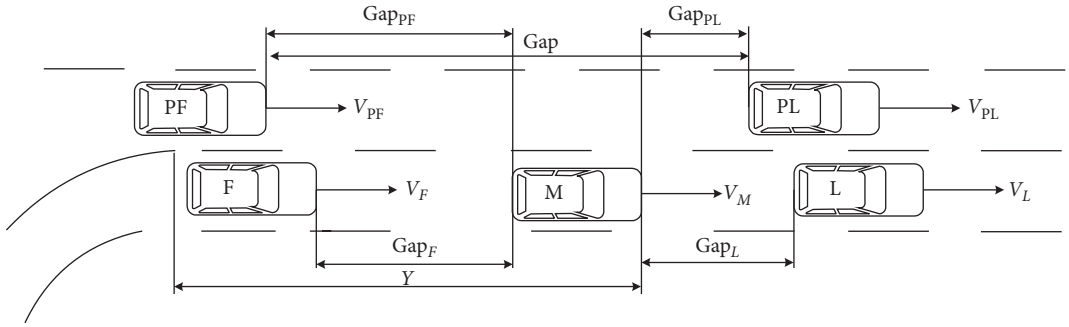

Figure 5: Schematic diagram of candidate variables.

Table 2: Candidate variables and explanations.

Candidate variables	Descriptions
V_M (m/s)	The speed of merging vehicle
V_{PL} (m/s)	The speed of putative leading vehicle
V_L (m/s)	The speed of putative following vehicle
ΔV_{PL} (m/s)	The relative speed between merging vehicle and its putative leading vehicle
ΔV_{PF} (m/s)	The relative speed between merging vehicle and its putative following vehicle
$\mathrm{Gap}_{\mathrm{PL}}$ (m)	The gap size between merging vehicle and its putative leading vehicle
$\mathrm{Gap}_{\mathrm{PF}}$ (m)	The gap size between merging vehicle and its putative following vehicle.
$(\mathrm{Gap}_{\mathrm{PF}}/\mathrm{Gap})$	The ratio of $\mathrm{Gap}_{\mathrm{PF}}$ to the total gap
V_L (m/s)	The speed of leading vehicle in the auxiliary lane
V_F (m/s)	The speed of following vehicle in the auxiliary lane
ΔV_L (m/s)	The speed difference between merging vehicle and its leading vehicle in the auxiliary lane
ΔV_F (m/s)	The speed difference between merging vehicle and its following vehicle in the auxiliary lane
Gap_L (m)	The gap size between merging vehicle and its leading vehicle
Gap_F (m)	The gap size between merging vehicle and its following vehicle
$\mathrm{TTC}_{\mathrm{PL}}$ (s)	The TTC between merging vehicle and putative leading vehicle
$\mathrm{TTC}_{\mathrm{PF}}$ (s)	The TTC between merging vehicle and putative following vehicle
TTC_L (s)	The TTC between merging vehicle and leading vehicle in the auxiliary lane
TTC_F (s)	The TTC between merging vehicle and following vehicle in the auxiliary lane
Y (m)	The longitudinal position of merging vehicle

leading vehicle, TTC will be infinite or too large. In order to restrict these situations, we will set the TTC range to $(0, 100\,\mathrm{s}]$, that is, when TTC is negative or greater than 100 s, it is set to 100 s.

Table 3 shows the main statistical characteristics of the candidate variables for merging behavior. One can find that the merging vehicles move faster than both PF and PL vehicles and the PF vehicles have the lowest average speed. Both the leading and following vehicles in the auxiliary lane move faster than the merging vehicles. Additionally, the average speed of merging vehicles reduces from 12.477 m/s to 12.086 m/s during the merging process to accommodate for the mainline traffic speed, which can also be reflected by changes of average ΔV_{PL} and ΔV_{PF}. It is interesting to find that $\mathrm{Gap}_{\mathrm{PF}}$ increases from 9.616 m to 16.081 m while $\mathrm{Gap}_{\mathrm{PL}}$ does not change much. It means $\mathrm{Gap}_{\mathrm{PF}}$ plays an important role and the PF vehicles tend to yield to the merging vehicles during the merging execution period. One can also find that the $\mathrm{TTC}_{\mathrm{PL}}$ has the lowest average value during the merging process, indicating that the traffic conflicts between the merging vehicles and PL vehicles might be the most serious.

A Pearson's correlation analysis is conducted to correlation coefficients between dependent variable and independent variables, as shown in Table 4. Bold values are the insignificant correlation coefficients at 0.95 confidence level. One can find that the dependent variable y_n^t has significant correlations with several independent variables, such as V_{PL} and $\mathrm{Gap}_{\mathrm{PF}}$. It is interesting to find that there is no significant correlation between $\mathrm{Gap}_{\mathrm{PL}}$ and y_n^t. $(\mathrm{Gap}_{\mathrm{PF}}/\mathrm{Gap})$ has the strongest correlation with y_n^t.

5. Modelling Results

After extracting enough data, the RF model is trained and tested in this section to verify the effectiveness. A data mining software called Salford Predictive Modeler is used in this study [16]. The data is randomly divided into two parts: 80% of the lane change cases are randomly selected as the training data, and the remaining 20% is used as the test data for validation. Though RF can use the OOB data for validation, we still do this for comparison with the state-of-the-art methods.

5.1. Parameter Determination. The number of decision trees B is an important parameter of RF. When building decision trees, RF does not prune it. Thus, the modeling

Table 3: Statistics of candidate influencing variables.

Candidate variables	$y_n^t = 0$				$y_n^t = 1$			
	Average	Standard deviation	Maximum	Minimum	Average	Standard deviation	Maximum	Minimum
V_M (m/s)	12.477	3.610	23.389	1.539	12.086	3.269	23.265	2.481
V_{PL} (m/s)	10.997	3.324	19.839	1.578	11.454	3.129	19.967	1.794
ΔV_{PL} (m/s)	1.480	2.233	13.089	−5.481	1.092	1.928	10.895	−5.247
Gap_{PL} (m)	13.699	16.781	172.256	0.491	13.821	15.507	152.789	0.631
V_{PF} (m/s)	10.320	3.157	18.681	0.501	10.912	2.925	18.868	1.923
ΔV_{PF} (m/s)	−2.157	2.247	4.344	−12.554	−1.175	1.845	4.113	−11.484
Gap_{PF} (m)	9.616	13.660	129.836	0.202	16.081	14.100	134.491	0.410
(Gap_{PF}/Gap)	0.316	0.274	1.241	0.001	0.452	0.1864	0.900	0.040
V_L (m/s)	14.864	3.541	23.543	3.661	15.550	3.106	21.909	6.610
ΔV_L (m/s)	−2.103	3.135	5.656	−13.708	−3.173	3.396	−15.255	7.837
Gap_L (m)	54.27	38.87	186.94	1.030	56.08	39.54	189.46	2.260
V_F (m/s)	13.282	3.161	21.730	2.585	13.764	3.108	21.566	2.387
ΔV_F (m/s)	0.980	3.010	13.445	−11.474	1.082	3.070	10.741	−12.470
Gap_F (m)	46.84	46.12	105.83	1.610	43.52	43.64	0.66	192.96
TTC_{PL} (s)	38.63	42.67	100	0.02	42.98	42.96	100	0.38
TTC_{PF} (s)	80.54	35.92	100	0.01	71.21	41.17	100	0.82
TTC_L (s)	85.94	29.09	100	0.36	83.07	30.99	100	0.42
TTC_F (s)	75.61	37.92	100	0.06	68.94	41.33	100	0.06
Y (m)	82.96	68.45	350.86	0.05	94.50	74.49	361.15	0.97

Table 4: Correlation coefficients between dependent variables and independent variables.

	Correlation	Coefficient P value
V_M	**−0.047**	**0.062**
V_{PL}	0.059	0.019
V_{PF}	0.081	0.001
ΔV_{PL}	−0.164	0.0001
ΔV_{PF}	0.190	0.0001
Gap_{PL}	**0.003**	**0.901**
Gap_{PF}	0.196	0.0001
(Gap_{PF}/Gap)	0.224	0.0001
V_L	0.084	0.013
V_F	−0.065	0.021
ΔV_L	−0.140	0.0001
ΔV_F	**0.014**	**0.618**
Gap_L	**0.020**	**0.564**
Gap_F	**−0.031**	**0.270**
TTC_{PL}	0.043	0.086
TTC_{PF}	−0.106	0.0001
TTC_L	**−0.043**	**0.210**
TTC_F	−0.076	0.007
Y	0.072	0.004

accuracy of RF will increase rapidly with the increase of the number of decision trees at first. However, after reaching a certain number, generating more trees would not improve the model accuracy but increase the computational burden. Previous studies showed that the total number of trees should be set at 200–500 [45, 50]. To ensure the reliability of the modeling results, this paper sets the number of trees at 500.

In RF, a randomly selected subset of features is used to build each single tree. Reducing the number of sampled features m would bring down the correlation among decision tree, leading to less generalization error. However, a too small m would also make the single tree suffer from large prediction error. Different m has been used in different studies [49, 58]; thus, the number of sampled features m should be selected carefully. To select the best m, RF models are trained with an increasing number of m from 1 to 10. Table 5 shows the OOB errors with a different number of m. One can find that the OOB error has the lowest value when m is 3. Thus, the number of randomly sampled features m is set at 3 in this study.

5.2. Variable Importance. The variable importance can be easily obtained by RF according to equation (2). The rank and importance values of independent variables are shown in Table 6.

Table 5: OOB errors with different m.

m	1	2	3	4	5	6	7	8	9	10
OOB error	9.6%	9.4%	9.1%	9.5%	9.5%	9.4%	9.7%	10.1%	10.4%	10.8%

Table 6: Rank of variable importance.

Rank	Variables	Importance value (%)
1	Gap_{PF}	27.35
2	$(\text{Gap}_{\text{PF}}/\text{Gap})$	23.33
3	Gap_{PL}	9.82
4	Y	8.68
5	ΔV_{PF}	6.82
6	TTC_{PL}	5.77
7	TTC_{PF}	3.69
8	ΔV_{PL}	3.46
9	ΔV_F	2.58
10	Gap_F	1.36
11	V_{PF}	1.31
12	V_F	1.28
13	V_M	1.28
14	V_L	1.03
15	ΔV_L	1.02
16	V_L	0.98
17	TTC_F	0.95
18	Gap_L	0.68
19	TTC_L	0.18

According to Table 6, it can be seen that Gap_{PF} and $(\text{Gap}_{\text{PF}}/\text{Gap})$ are the most two important variables, whose importance values are much greater than other variables. The reason is probably that merging vehicle drivers can easily observe the PL vehicles and control the relative speeds and positions with them. Thus, they tend to leave more space for their PF vehicles. This finding is consistent with that of the previous studies [20].

5.3. Feature Variable Selection. From Table 6, one can find that the relative importance values of several variables are rather low, such as TTC_L (0.18%), indicating that there are some redundant or irrelevant variables in the RF model. Therefore, a feature variable selection process introduced by Genuer et al. [59] is applied in this study. The basic steps are shown as follows:

(1) Build a RF model with all candidate variables and rank the variables with the relative importance values in descending order

(2) Delete the variable with the lowest relative importance value and create a new variable set

(3) Build a new RF model with the new variable set and rank the variables with the relative importance values in descending order

(4) Repeat steps (2) and (3) until only one variable remains

(5) Rank all the RF models established in steps (1) to (4) according to the OOB error, and select the model and feature variable set with the lowest error

After feature variable selection, nine feature variables are remained and the OOB error is reduced from 9.1% to 8.9%, indicating that reducing the number of feature variables will not reduce the prediction performance. The values of variable importance in the model are shown in Table 7. It is easy to know from Table 7 that Gap_{PF} and $(\text{Gap}_{\text{PL}}/\text{Gap})$ are still the two most important factors. ΔV_F is the only variable related to the vehicles in the auxiliary lane, which means merging vehicle drivers mainly focus on the traffic condition in the main lane.

5.4. Accuracy of the Model. Table 8 shows the prediction accuracy for training data and testing data. For comparison, a binary logit model and a CART model are also built based on the same dataset. Significant variables are selected by stepwise selection method. The final binary logit model is shown as

$$P\left(y_n^t\right) = \frac{1}{1 + \exp\left(1.710 - 0.0829\Delta V_{\text{PL}} - 0.1481\Delta V_{\text{PF}} + 0.1321\Delta V_L - 0.01551\text{Gap}_{\text{PL}} - 2.076\left(\text{Gap}_{\text{PF}}/\text{Gap}\right) - 0.0405Y\right)}. \quad (5)$$

Table 7: Rank of variable importance after variable selection.

Rank	Selected variables	Importance values (%)
1	Gap_{PF}	30.59
2	(Gap_{PF}/Gap)	27.05
3	Gap_{PL}	12.99
4	Y	6.66
5	ΔV_{PF}	7.32
6	TTC_{PL}	7.14
7	TTC_{PF}	5.58
8	ΔV_{PL}	5.01
9	ΔV_F	4.31

Table 8: Prediction results and comparison of models.

Data set	Random forest (%)	Binary logit model (%)	CART model (%)
Training data	91.1	78.9	95.4
Testing data	88.3	72.5	76.29

The results show that the prediction accuracy of the RF model is much better than the binary logit model for both training data and test data. One can also find that CART has the highest prediction accuracy in training data. However, the performance of CART in testing data is much poorer than RF, indicating that RF has better ability to deal with problem of overfitting than CART. In addition, due to the influence of collinearity of variables, only six variables are included in the binary logit model. Some variables that may affect the merging decision behavior in a certain range are ignored by the binary logit model, such as TTC_{PL} and ΔV_F. It is clear that RF can overcome the collinearity problem and deeply explore the complicated nonlinear relationships between merging decision and influencing variables. One can also find that the reduction of the accuracy in training and testing dataset is also much smaller than the logit model and CART model, showing that RF is practical for predicting the merging decision during execution period and has better transferability.

6. Conclusions

This study conducts a comprehensive analysis of the influencing variables of merging decision and employs the random forest (RF) to model the merging decision behavior during the execution period. The proposed RF method can accurately predict the merging decision during the execution period and investigate important influencing factors. The US-101 vehicle trajectory data are used to train and validate the RF model. To comprehensively explore the influencing factors during merging execution, 19 candidate variables are extracted including speeds, relative speeds, gaps, time-to-collisions (TTCs), and locations.

The modeling results show that Gap_{PF} and (Gap_{PF}/Gap) are the most two important variables, whose importance values are much greater than other variables. It is probably because that the merging vehicle drivers can easily observe the PL vehicles and control the relative speeds and positions with them and thus, they tend to leave more space for their PF vehicles. To select the effective variables, a feature variable selection process is adopted and 9 variables are selected in the RF model finally. Gap_{PF} and (Gap_{PF}/Gap) are still the two most important feature variables. ΔV_F is the only variable related to the vehicles in the auxiliary lane, which means merging vehicles mainly focus on the traffic condition in the adjacent main lane. Evaluation of the performances in comparison with the state-of-the-art method reveals that the proposed method can obtain much more accurate results in both training ant testing datasets. The reduction of the accuracy in training and testing dataset is also much smaller than that of logit model, showing that RF is practical for predicting the merging decision behavior during execution period and has better transferability.

Furthermore, it is obvious that merging drivers face more challenges and may make improper decisions under congested traffic conditions, which might cause long delays. In future, if vehicles can receive the real-time information about the traffic environment via VANETs, the proposed RF models can help the merging vehicles make safer decisions. Thus, the results of this study can also improve the safety and comfort of driving assistance systems and autonomous driving systems.

Data Availability

The NGISM data used to support the findings of this study have been deposited at the website: https://catalog.data.gov/dataset/next-generation-simulation-ngsim-vehicle-trajectories.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the Science and Technology Innovation Fund for Youth Scientists of Nanjing Forestry University, grant number CX2019021, and Scientific Research Start-up Funds of Nanjing Forestry University, grant number GXL2020012.

References

[1] L. Elefteriadou, R. P. Roess, and W. R. McShane, "Probabilistic nature of breakdown at freeway merge junctions," *Transportation Research Record*, vol. 1995, no. 1484, pp. 80–89, 1995.

[2] M. Li, Z. Li, C. Xu, and T. Liu, "Short-term prediction of safety and operation impacts of lane changes in oscillations with empirical vehicle trajectories," *Accident Analysis & Prevention*, vol. 135, Article ID 105345, 2020.

[3] X. Gu, M. Abdel-Aty, Q. Xiang, Q. Cai, and J. Yuan, "Utilizing UAV video data for in-depth analysis of drivers' crash risk at interchange merging areas," *Accident Analysis & Prevention*, vol. 123, pp. 159–169, 2019.

[4] C. Yin, J. Zhang, C. Shao, and Society, "Relationships of the multi-scale built environment with active commuting, body mass index, and life satisfaction in China: a GSEM-based

analysis," *Travel Behaviour and Society*, vol. 21, pp. 69–78, 2020.

[5] E. C. Olsen, S. E. Lee, W. W. Wierwille, and M. J. Goodman, "Analysis of distribution, frequency, and duration of naturalistic lane changes," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 46, no. 22, pp. 1789–1793, 2002.

[6] Z. Yao, L. Shen, R. Liu, Y. Jiang, and X. Yang, "A dynamic predictive traffic signal control framework in a cross-sectional vehicle infrastructure integration environment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 4, pp. 1455–1466, 2019.

[7] Z. Liu, Y. Liu, Q. Meng, and Q. Cheng, "A tailored machine learning approach for urban transport network flow estimation," *Transportation Research Part C: Emerging Technologies*, vol. 108, pp. 130–150, 2019.

[8] Z. Yao, T. Xu, Y. Jiang, and R. Hu, "Linear stability analysis of heterogeneous traffic flow considering degradations of connected automated vehicles and reaction time," *Physica A: Statistical Mechanics and its Applications*, vol. 561, Article ID 125218, 2021.

[9] Z. Yao, R. Hu, Y. Jiang, and T. Xu, "Stability and safety evaluation of mixed traffic flow with connected automated vehicles on expressways," *Journal of Safety Research*, vol. 75, pp. 262–274, 2020.

[10] H. Wang, Y. Qin, W. Wang, and J. Chen, "Stability of CACC-manual heterogeneous vehicular flow with partial CACC performance degrading," in *Transportmetrica B: Transport Dynamics*, vol. 7, no. 1, pp. 788–813, 2019.

[11] J. Weng, S. Xue, and X. Yan, "Modeling vehicle merging behavior in work zone merging areas during the merging implementation period," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 917–925, 2016.

[12] M. Ardakani, J. Yang, and L. Sun, "Stimulus response driving behavior: an improved general motor vehicle-following model," *Advances in Transportation Studies*, vol. 39, 2016.

[13] M. K. Ardakani and J. Yang, "Generalized Gipps-type vehicle-following models," *Journal of Transportation Engineering*, vol. 143, no. 3, pp. 04016011.1–04016011.10, 2017.

[14] V. Alexiadis, J. Colyar, J. Halkias, R. Hranac, and G. McHale, "The next generation simulation program," *Institute of Transportation Engineers ITE Journal*, vol. 74, no. 8, pp. 22–26, 2004.

[15] G. Li and J. Cheng, "Exploring the effects of traffic density on merging behavior," *IEEE Access*, vol. 7, pp. 51608–51619, 2019.

[16] G. Li, S. Fang, J. Ma, and J. Cheng, "Modeling merging acceleration and deceleration behavior based on gradient-boosting decision tree," *Journal of Transportation Engineering*, vol. 146, no. 7, Article ID 05020005, 2020.

[17] G. Li and L. Sun, "Characterizing heterogeneity in drivers' merging maneuvers using two-step cluster analysis," *Journal of Advanced Transportation*, vol. 2018, Article ID 5604375, 2018.

[18] X. Wan, P. J. Jin, H. Gu, X. Chen, and B. Ran, "Modeling freeway merging in a weaving section as a sequential decision-making process," *Journal of Transportation Engineering, Part A: Systems*, vol. 143, no. 5, Article ID 05017002, 2017.

[19] K. I. Ahmed, *Modeling Drivers' Acceleration and Lane Changing Behavior*, Massachusetts Institute of Technology, Cambridge, MA, USA, 1999.

[20] C. F. Choudhury, V. Ramanujam, and M. E. Ben-Akiva, "Modeling acceleration decisions for freeway merges," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2124, no. 1, pp. 45–57, 2009.

[21] G. Li, Y. Pan, Z. Yang, and J. Ma, "Modeling vehicle merging position selection behaviors based on a finite mixture of linear regression models," *IEEE Access*, vol. 7, pp. 158445–158458, 2019.

[22] E. Balal, R. L. Cheu, T. Gyan-Sarkodie, and J. Miramontes, "Analysis of discretionary lane changing parameters on freeways," *International Journal of Transportation Science and Technology*, vol. 3, no. 3, pp. 277–296, 2014.

[23] P. G. Gipps, "A model for the structure of lane-changing decisions," *Transportation Research Part B: Methodological*, vol. 20, no. 5, pp. 403–414, 1986.

[24] Q. Yang and H. N. Koutsopoulos, "A microscopic traffic simulator for evaluation of dynamic traffic management systems," *Transportation Research Part C: Emerging Technologies*, vol. 4, no. 3, pp. 113–129, 1996.

[25] L. Wu, L. Yu, W. Wang, and J. Liu, "Prediction for the region disposition of Panama dry bulk fleet management," *IEEE Access*, vol. 7, pp. 136604–136615, 2019.

[26] P. Hidas, "Modelling vehicle interactions in microscopic simulation of merging and weaving," *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 1, pp. 37–62, 2005.

[27] L. Bloomberg and J. Dale, "A comparison of the VISSIM and CORSIM traffic simulation models on a congested network," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1727, no. 1, pp. 52–60, 2000.

[28] J. Weng and Q. Meng, "Modeling speed-flow relationship and merging behavior in work zone merging areas," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 985–996, 2011.

[29] F. Marczak, W. Daamen, and C. Buisson, "Merging behaviour: empirical comparison between two sites and new theory development," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 530–546, 2013.

[30] J. Weng, S. Xue, Y. Yang, X. Yan, and X. Qu, "In-depth analysis of drivers' merging behavior and rear-end crash risks in work zone merging areas," *Accident Analysis & Prevention*, vol. 77, pp. 51–61, 2015.

[31] G. Li, "Application of finite mixture of logistic regression for heterogeneous merging behavior analysis," *Journal of Advanced Transportation*, vol. 2018, Article ID 1436521, 2018.

[32] D. Arbis and V. V. Dixit, "Game theoretic model for lane changing: incorporating conflict risks," *Accident Analysis & Prevention*, vol. 125, pp. 158–164, 2019.

[33] K. Kang and H. A. Rakha, "Modeling driver merging behavior: a repeated game theoretical approach," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 20, pp. 144–153, 2018.

[34] Q. Meng and J. Weng, "Classification and regression tree approach for predicting drivers' merging behavior in short-term work zone merging areas," *Journal of Transportation Engineering*, vol. 138, no. 8, pp. 1062–1070, 2012.

[35] E. Balal, R. L. Cheu, and T. Sarkodie-Gyan, "A binary decision model for discretionary lane changing move based on fuzzy inference system," *Transportation Research Part C: Emerging Technologies*, vol. 67, pp. 47–61, 2016.

[36] J. Tang, F. Liu, W. Zhang, R. Ke, and Y. Zou, "Lane-changes prediction based on adaptive fuzzy neural network," *Expert Systems with Applications*, vol. 91, pp. 452–463, 2018.

[37] Y. Hou, P. Edara, and C. Sun, "Modeling mandatory lane changing using Bayes classifier and decision trees," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 647–655, 2014.

[38] E. Wang and J. Sun, "Exploring freeway merging behavior using dynamic bayesian network models," in *Proceedings of*

the International Conference on Transportation and Development 2018: Traffic and Freight Operations and Rail and Public Transit, pp. 120–130, American Society of Civil Engineers, Reston, VA, USA, July 2018.

[39] S. Moridpour, M. Sarvi, G. Rose, and E. Mazloumi, "Lane-changing decision model for heavy vehicle drivers," *Journal of Intelligent Transportation Systems*, vol. 16, no. 1, pp. 24–35, 2012.

[40] L. Xu, J. Lu, B. Ran, F. Yang, and J. Zhang, "Cooperative merging strategy for connected vehicles at highway on-ramps," *Journal of Transportation Engineering, Part A: Systems*, vol. 145, no. 6, Article ID 04019022, 2019.

[41] M. Vasconcelos and N Vasconcelos, "Natural image statistics and low-complexity feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 228–244, 2008.

[42] H. Yang and J. Moody, "Data visualization and feature selection: new algorithms for nongaussian data," *Neural Information Processing Systems*, vol. 12, pp. 687–693, 1999.

[43] T. Ye, C. Zu, B. Jie, D. Shen, and D. Zhang, "Discriminative multi-task feature selection for multi-modality classification of Alzheimer's disease," *Brain Imaging and Behavior*, vol. 10, no. 3, pp. 739–749, 2016.

[44] W. Shu and H. Shen, "Incremental feature selection based on rough set in dynamic incomplete data," *Pattern Recognition*, vol. 47, no. 12, pp. 3890–3906, 2014.

[45] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[46] W. Tang and D. M. Levinson, "Deviation between actual and shortest travel time paths for commuters," *Journal of Transportation Engineering, Part A: Systems*, vol. 144, no. 8, Article ID 04018042, 2018.

[47] M. Belgiu, L. Drăguţ, and R. Sensing, "Random forest in remote sensing: a review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.

[48] B. Hamner, "Predicting travel times with context-dependent random forests by modeling local and aggregate traffic flow," in *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, pp. 1357–1359, IEEE, Sydney, Australia, December 2010.

[49] L. Cheng, X. Chen, J. De Vos, X. Lai, and F. Witlox, "Applying a random forest method approach to model travel mode choice behavior," *Travel Behaviour and Society*, vol. 14, pp. 1–10, 2019.

[50] J. Cheng, G. Li, and X. Chen, "Developing a travel time estimation method of freeway based on floating car using random forests," *Journal of Advanced Transportation*, vol. 2019, Article ID 8582761, 2019.

[51] Y. Hou, P. Edara, and Y. Chang, "Road network state estimation using random forest ensemble learning," in *Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–6, IEEE, Yokohama, Japan, October 2017.

[52] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 308–324, 2015.

[53] Q. Wang, Z. Li, and L. Li, "Investigation of discretionary lane-change characteristics using next-generation simulation data sets," *Journal of Intelligent Transportation Systems*, vol. 18, no. 3, pp. 246–253, 2014.

[54] X. Wan, P. J. Jin, F. Yang, and B. Ran, "Merging preparation behavior of drivers: how they choose and approach their merge positions at a congested weaving area," *Journal of Transportation Engineering*, vol. 142, no. 9, Article ID 05016005, 2016.

[55] V. Punzo, M. T. Borzacchiello, and B. Ciuffo, "On the assessment of vehicle trajectory data accuracy and application to the next generation SIMulation (NGSIM) program data," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 1243–1262, 2011.

[56] Cambridge Systematics, "NGSIM US 101 data analysis: summary report," in *Prepared for Federal Highway Administration*Cambridge Systematics, Cambridge, UK, 2005.

[57] C. Thiemann, M. Treiber, and A. Kesting, "Estimating acceleration and lane-changing dynamics from next generation simulation trajectory data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2008, pp. 90–101, 2008.

[58] M. Ghasri, T. Hossein Rashidi, and S. T. Waller, "Developing a disaggregate travel demand system of models using data mining techniques," *Transportation Research Part A: Policy and Practice*, vol. 105, pp. 138–153, 2017.

[59] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.